

E5873

(19)日本国特許庁(JP)

(12) 公開特許公報(A)

(11)特許出願公開番号

特開平6-139027

(43)公開日 平成6年(1994)5月20日

(51)Int.Cl.⁵

識別記号

庁内整理番号

FI

技術表示箇所

G 0 6 F 3/06

3 0 1 Z

7165-5B

3 0 4 F

7165-5B

G 1 1 B 19/02

F

7525-5D

審査請求 未請求 請求項の数20(全 22 頁)

(21)出願番号

特願平4-290428

(22)出願日

平成4年(1992)10月28日

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72)発明者 大枝 高

神奈川県横浜市戸塚区吉田町292番地 株

式会社日立製作所マイクロエレクトロニク

ス機器開発研究所内

(72)発明者 吉田 稔

神奈川県小田原市国府津2880番地 株式会

社日立製作所小田原工場内

(74)代理人 弁理士 武 顕次郎

最終頁に続く

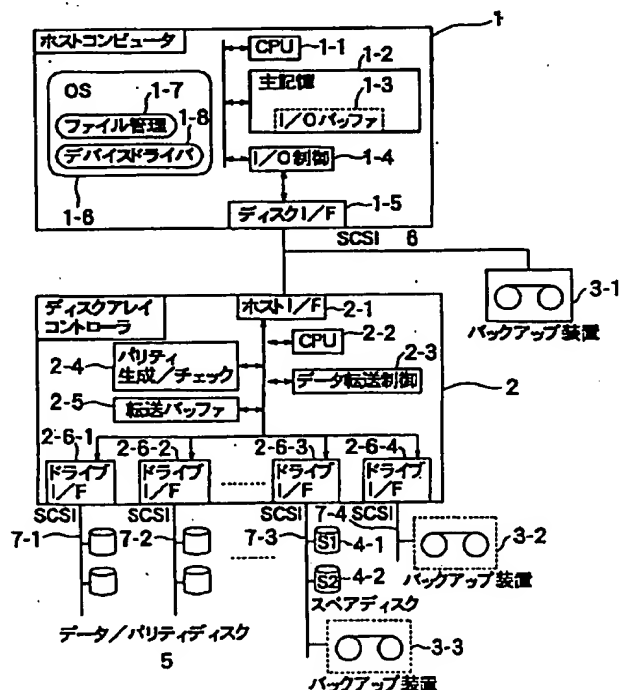
(54)【発明の名称】 ディスクドライバ、ディスクアレイ装置、データ記憶システム及びディスクアレイシステムのデータバックアップ方法

(57)【要約】

【目的】 ディスクアレイ装置のデータのバックアップ占有時間を短縮する。

【構成】 ディスクアレイコントローラ2に、ドライブI/F 2-6-1, 2-6-2, ……を介し、データやパリティ情報を格納するデータ/パリティディスクドライブ5がアレイ状に接続されており、また、このディスクアレイコントローラ2のドライブI/F 2-6-3にスベアディスクドライブ4が所定個数接続されている。かかるスベアディスクドライブ4は、データ/パリティディスクドライブ5に障害が生じたとき、この障害データ/パリティディスクドライブに代ってデータ/パリティディスクドライブとなるが、かかるディスクアレイ上でのデータをバックアップ装置3-1, 3-2または3-3でバックアップのためのデータ転送に際し、データ/パリティディスクドライブ5とかかるバックアップ装置との間のデータ転送バッファにもなる。

【図 1】



E5873

MENU

SEARCH

INDEX

DETAIL

1 / 1

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 06-139027

(43)Date of publication of application : 20.05.1994

(51)Int.Cl. G06F 3/06
G06F 3/06
G11B 19/02

(21)Application number : 04-290428

(71)Applicant : HITACHI LTD

(22)Date of filing : 28.10.1992

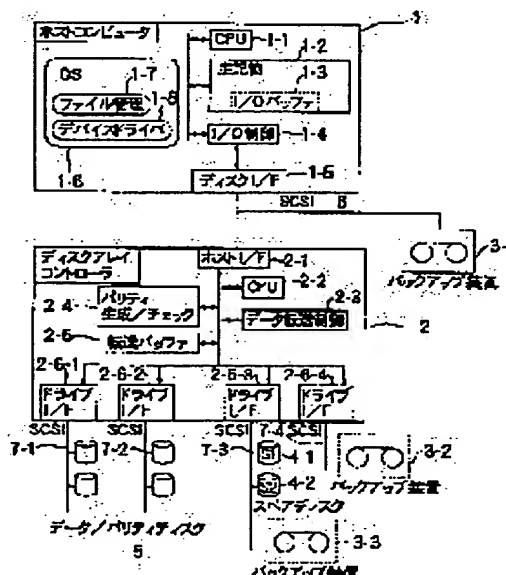
(72)Inventor : OEDA TAKASHI
YOSHIDA MINORU
HONDA KIYOSHI
MATSUNAMI NAOTO
MIYAZAWA SHOICHI
ISONO SOICHI

(54) DATA BACKUP METHOD FOR DISK DRIVER AND DISK ARRAY DEVICE AND
DATA STORAGE SYSTEM AND DISK ARRAY SYSTEM

(57)Abstract:

PURPOSE: To shorten the backup occupying time of a disk array device.

CONSTITUTION: A data/parity disk drive 5 which stores data or parity information is connected like an array through drive I/F 2-6-1, 2-6-2,... with a disk array controller 2. And also, the prescribed number of spare disk drives 4 are connected with the drive I/F 2-6-3 of the disk array controller 2. When a fault occurs at the data/parity disk drive 5, the spare disk drive 4 can be used as the data/parity disk drive instead of the defective data parity/parity disk drive, or used as a data transfer buffer between the data/parity disk drive 5 and backup devices 3-1, 3-2, and 3-3 at the time of a data transfer for backup- processing data on the disk array by backup devices.



LEGAL STATUS

[Date of request for examination] 11.03.1997

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

【特許請求の範囲】

【請求項1】 ディスクアレイコントローラと、ドライ
ブインターフェースによって該ディスクアレイコント
ローラに接続され、アレイ状に配列された複数のデータ等
を格納するためのディスクドライブと、該ディスクドラ
イブに格納されているデータをバックアップするための
データバックアップ装置とを備え、

該複数のディスクドライブのうちの1以上のディスクド
ライブを予備ディスクドライブとし、残りをデータ等を
格納するデータ/パリティディスクドライブとし、該予
備ディスクドライブを、障害が生じた該データ/パリテ
ィディスクドライブに代ってデータ/パリティディス
クドライブとするディスクアレイ装置において、
データのバックアップの際、該予備ディスクドライブを
該データ/パリティディスクドライブと該データバック
アップ装置との間のデータ転送バッファとすることを特
徴とするディスクアレイ装置。

【請求項2】 請求項1において、
前記データバックアップ装置を、前記ディスクアレイコ
ントローラのドライブインターフェースに接続し、前記
ディスクアレイコントローラからの制御を可能に構成し
たことを特徴とするディスクアレイ装置。

【請求項3】 請求項2において、
前記データ/パリティディスクドライブはインターフェ
ースを2ポート以上有し、前記ディスクアレイコント
ローラと前記データバックアップ装置とに接続されてい
ることを特徴とするディスクアレイ装置。

【請求項4】 請求項1において、
前記アレイ上に配列された前記複数のディスクドライブ
上での現在のファイルシステムの使用範囲を認識する手
段もしくは上位コンピュータから、その認識結果を表わ
す情報を獲得する手段を有し、
該上位コンピュータが介在することなく、該ファイルシ
ステムの使用範囲のデータの前記データバックアップ装
置へのバックアップの進行、及び前記アレイ上に配列さ
れた前記複数のディスクドライブ上への前記データバック
アップ装置のバックアップデータの回復を行なうこと
を特徴とするディスクアレイ装置。

【請求項5】 請求項4において、
前回のバックアップ時から変更されたアドレスを認識す
る手段もしくは前記上位コンピュータから、その認識結
果を表わす情報を獲得する手段を有し、
前記上位コンピュータが介在することなく、前記データ
バックアップ装置への差分バックアップの進行、及び前
記アレイ上に配列された前記複数のディスクドライブ上
への前記データバックアップ装置のバックアップデータ
の回復を行なうことを特徴とするディスクアレイ装置。

【請求項6】 請求項1において、
前記データバックアップ装置へのバックアップが実行さ
れている前記アレイ上に配列された前記複数のディス

クドライブ上でのバックアップ対象ファイルシステムへの
アクセスを記録しておく手段を有し、

上位コンピュータから該バックアップ対象ファイルシ
ステムへのアクセスをしながら、該バックアップ対象フ
ァイルシステムのデータのバックアップができることを特
徴とするディスクアレイ装置。

【請求項7】 1もしくは複数のインターフェースによ
ってコンピュータシステムと接続され、かつ1もしくは
複数のドライブインターフェース夫々に複数の記憶装置
が接続されてなり、該複数の記憶装置によってファイル
データを記憶できるようにしたデータ記憶システムにお
いて、
該記憶装置の少なくとも1つは予備の記憶装置であっ
て、

データバックアップ手段と、
データを記憶する記憶装置に障害が発生したとき、該障
害が発生した記憶装置に代えて、その論理的位置を該予
備の記憶装置で占めさせる手段と、
該記憶装置のデータのデータバックアップ時、該予備の
記憶装置を該記憶装置と該バックアップ手段との間のデ
ータ転送バッファとする手段とを有することを特徴とす
るデータ記憶システム。

【請求項8】 請求項7において、
前記複数の記憶装置上での現在のファイルシステムの使
用範囲を認識する手段もしくは上位コンピュータから、
その認識結果を表わす情報を獲得する手段を有し、
該上位コンピュータが介在することなく、該ファイルシ
ステムの使用範囲のデータの前記データバックアップ手
段へのバックアップの進行、及び前記複数の記憶装置上
への前記データバックアップ手段のバックアップデータ
の回復を行なうことを特徴とするデータ記憶システム。

【請求項9】 請求項8において、
前回のバックアップ時から変更されたアドレスを認識す
る手段もしくは前記上位コンピュータから、その認識結
果を表わす情報を獲得する手段を有し、
前記上位コンピュータが介在することなく、前記データ
バックアップ手段への差分バックアップの進行、及び前
記複数の記憶装置上への前記データバックアップ手段の
バックアップデータの回復を行なうことを特徴とするデ
ータ記憶システム。

【請求項10】 請求項7において、
前記データバックアップ手段へのバックアップが実行さ
れている前記複数の記憶装置上でのバックアップ対象フ
ァイルシステムへのアクセスを記録しておく手段を有
し、
上位コンピュータから該バックアップ対象ファイルシ
ステムへのアクセスをしながら、該バックアップ対象フ
ァイルシステムのデータのバックアップができることを特
徴とするデータ記憶システム。

【請求項11】 ディスクアレイコントローラと、ドラ

イブインターフェースによって該ディスクアレイトローラに接続され、アレイトに配列された複数のデータ等を格納するためのディスクドライブと、該ディスクドライブに格納されているデータをバックアップするためのデータバックアップ装置とを備え、

該複数のディスクドライブのうちの1以上のディスクドライブを予備ディスクドライブとし、残りをデータ等を格納するデータ/パリティディスクドライブとし、該予備ディスクドライブを、障害が生じた該データ/パリティディスクドライブに代ってデータ/パリティディスクドライブとするディスクアレイトシステムにおいて、データのバックアップの際、該予備ディスクドライブを該データ/パリティディスクドライブと該データバックアップ装置との間のデータ転送バッファとすることを特徴とするディスクアレイトシステムのデータバックアップ方法。

【請求項12】 請求項11において、前記データバックアップ装置は前記ディスクアレイトローラによって制御されることを特徴とするディスクアレイトシステムのデータバックアップ方法。

【請求項13】 請求項11において、前記ディスクドライブはインターフェースを2ポート以上有し、前記ディスクアレイトローラと前記データバックアップ装置とに接続されていることを特徴とするディスクアレイトシステムのデータバックアップ方法。

【請求項14】 請求項11において、前記アレイト上に配列された前記複数のディスクドライブ上での現在のファイルシステムの使用範囲を認識する手段もしくは上位コンピュータから、その認識結果を表わす情報を獲得し、該上位コンピュータが介在することなく、該ファイルシステムの使用範囲のデータの前記データバックアップ装置へのバックアップの進行、及び前記アレイト上に配列された前記複数のディスクドライブ上への前記データバックアップ装置のバックアップデータの回復を行なうことを特徴とするディスクアレイトシステムのデータバックアップ方法。

【請求項15】 請求項14において、前回のバックアップ時から変更されたアドレスを認識する手段もしくは前記上位コンピュータから、その認識結果を表わす情報を獲得し、前記上位コンピュータが介在することなく、前記データバックアップ装置への差分バックアップの進行、及び前記アレイト上に配列された前記複数のディスクドライブ上への前記データバックアップ装置のバックアップデータの回復を行なうことを特徴とするディスクアレイトシステムのデータバックアップ方法。

【請求項16】 請求項14において、前記データバックアップ装置へのバックアップが実行されている前記アレイト上に配列された前記複数のディスク

ドライブ上でのバックアップ対象ファイルシステムへのアクセスを記録しておき、

上位コンピュータから該バックアップ対象ファイルシステムへのアクセスをしながら、該バックアップ対象ファイルシステムのデータのバックアップができることを特徴とするディスクアレイトシステムのデータバックアップ方法。

【請求項17】 ディスクから読み出したデータを一時的に格納する半導体メモリを有するディスクドライブにおいて、

該ディスクの記録領域の所定部分を書込み禁止領域に設定する第1の手段と、

該半導体メモリの記録領域を複数に分割して管理する第2の手段とを備え、

該半導体メモリにおける該第2の手段によって分割形成された1乃至複数の記憶領域に、該第1の手段によって設定された書込み禁止領域の一部または全部から読み出したデータを記録して退避させ、しかる後に、該ディスクでの書込み禁止領域とした記録領域を書込み可能とすることを特徴とするディスクドライブ。

【請求項18】 請求項17において、ホスト装置と接続する手段を複数設けたことを特徴とするディスクドライブ。

【請求項19】 請求項17または18に記載のディスクドライブが複数個とアレイト制御装置とからなディスクアレイト装置において、

該ディスクドライブのデータバックアップの際、該アレイト制御装置の制御のもとに、データバックアップ領域に対する前記書込み禁止領域の設定と前記半導体メモリへのデータ退避動作とを行なうことを特徴とするディスクアレイト装置。

【請求項20】 請求項19において、前記ディスクドライブと前記アレイト制御装置とを接続する経路を複数個設け、前記書込み禁止領域の設定と退避動作との命令を伝達する前記接続経路とバックアップデータを前記ディスク装置の設定された書込み禁止領域から前記半導体メモリに転送する前記接続経路とを別にすることを特徴とするディスクアレイト装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】 本発明は、ディスクドライブ等のデータの記憶装置を所定個数備えたコンピュータシステムに係り、特に、該記憶装置に記憶されているデータのバックアップに関する。

【0002】

【従来の技術】 ディスクアレイト装置については、例えば特開平1-250128号公報に開示されている。この装置の特徴は、複数のディスクドライブが同時多重動作することによる高速化と、冗長ディスクドライブや予備

ディスクドライブを持つことによる高信頼化とを実現したことである。

【0003】ディスクドライブを高速化するためには、データ転送速度の高速化及び1秒当りのリード／ライト処理件数（I/Oスループット）の向上とを図ることが必要であるが、ディスクアレイ装置では、複数台のディスクドライブに対して並列にデータ転送を可能とすることにより、データ転送速度の高速化を実現している。かかる構成に、信頼度を上げるため、後述する冗長度を付加した構成を、一般に、RAID3と呼んでいる。また、I/Oスループットを向上させるために、複数台のディスクドライブに多重シーク／回転待ちさせるようにした構成があり、かかる構成に冗長度を付加して信頼度を上げるようにした構成を、RAID5と呼んでいる。

【0004】さて、以上のような性能を向上させるためには、上記いずれの構成においても、複数台のディスクドライブが必要となる。現在の標準的なディスクドライブの平均故障時間（MTBF）は5万時間程度であるが、このようにN台のディスクドライブを用いたディスクアレイ装置での平均故障時間（MTBF）は、5万／N時間に抑えられる。しかし、この値は、標準的な10台構成のディスクアレイ装置の場合5千時間、即ち七ヶ月弱になり、システムとしては、一般に、許容し難い値

【数1】

$$MTBF = \frac{(MTBF)^2}{N \cdot G \cdot (G+1) \cdot MTTR}$$

MTBF … ディスクアレイ装置全体の平均故障時間

MTBF … ディスクアレイ装置に用いられるディスク単体の平均故障時間

MTTR … ディスクの平均修理時間

N … パリティグループの数

G … パリティグループ中のデータディスク数

【0008】この数1によると、障害ディスクドライブを正常なディスクドライブと交換し、このディスクドライブへデータを回復するまでの時間、即ちディスクドライブの平均修理時間（MTTR）が平均故障時間（MTBF）を決定する重要なファクターになっていることが判る。この平均修理時間（MTTR）を最小限にするためには、障害ディスクドライブと交換する予備ディスクドライブを予めパリティグループ内に備えておき、障害が発生したとき、速やかに予備ディスクドライブを用いてデータの回復を行なうことが効果的である。かかる予備ディスクドライブについては、スベアディスクドライブとして、前記特開平1-250128号公報にも開示されている。

である。

【0005】この問題を解決するために、ディスクアレイ装置のディスクドライブを数台のデータディスクドライブと1台のパリティディスクドライブとからなるグループに区分し、また、データをストライプと呼ばれるデータ単位に分割し、上記グループ内の全ディスクドライブでの同じ論理アドレスから開始する全ストライプでパリティグループと呼ばれるグループを構成して、このパリティグループ内の全データストライプの排他的論理和を計算することによってパリティ情報を生成し、これをそのパリティグループのパリティディスクドライブにパリティストライプとして保存するようにした構成が知られている。かかる構成はRAIDと呼ばれるものであって、単にデータを複数のディスクドライブに分配するだけでなく、パリティグループ中の任意の1台が故障しても、パリティディスクドライブに保存されているパリティ情報から故障ディスクドライブのデータを回復することができるものであって、信頼性が向上する。

【0006】ところで、RAIDの平均故障時間（MTBF）は次の数1で表わされる。

【0007】

【数1】

【0009】RAID構成の採用とスベアディスクドライブの採用とにより、複数台のディスクドライブからなるディスクアレイ装置での平均故障時間（MTBF）の低下の問題は解決される。

【0010】

【発明が解決しようとする課題】しかし、上記数1で表わされる平均故障時間（MTBF）は、システム内の複数のディスクドライブの障害発生という事象が互いに独立であるという仮定をもとにしたものであり、地震・火災等の大規模災害の場合や、障害回復作業中のミスによる2重障害の発生の場合等は考慮されていない。また、オペレーションのミスによるデータの消失についても考慮されていない。

【0011】かかる問題を解消するためには、データのバックアップを定期的に行なうこと以外に現実的方法はなく、上記従来技術では、ディスクドライブのアレイ化による容量の増大に対応した効果的なバックアップ方法については、配慮されていない。実際、標準的な5.25" ディスクドライブ(1.5Gバイト)を使用し、かかるディスクドライブ10台でバリティグループを構成してそのうちの1台をスペアディスクドライブに、他の1台をバリティディスクドライブに夫々に割り当てた場合、このバリティグループの総容量は10GBとなり、標準的なバックアップ用テープドライブ(例えば、100kバイト/secあるいは約200kバイト/sec)でバックアップするのに要する時間は14時間弱である。これでは、ディスクアレイ装置の高I/Oスループットを生かしたオンライントランザクションシステムに応用する場合、大きな障害となる。毎日データのバックアップを行なう場合、この14時間、システムのサービスを停止してバックアップ作業を行わなければならないからである。

【0012】本発明の目的は、かかる問題を解消し、データバックアップの占有時間を大幅に短縮して、かつデータバックアップを効果的に行なうことができるようにしたディスクドライブ、ディスクアレイ装置、データ記憶システム及びディスクアレイシステムのデータバックアップ方法を提供することにある。

【0013】また、本発明の他の目的は、システムのサービスを閉塞せず、有効なデータバックアップを行なうことができるようにしたディスクドライブ、ディスクアレイ装置、データ記憶システム及びディスクアレイシステムのデータバックアップ方法を提供することにある。

【0014】

【課題を解決するための手段】上記目的を達成するためには、本発明は、データが記憶するための情報記憶装置に障害が発生したとき、この障害情報記憶装置に代えてデータを記憶するための情報記憶装置とする予備の記憶装置を、該データのバックアップ時、該情報記憶装置とデータバックアップ装置との間のデータ転送バッファとする。

【0015】また、本発明は、前記データバックアップ装置でバックアップされている前記情報記憶装置上でのバックアップ対象ファイルシステムへのアクセス情報を記録しておく手段を有し、該アクセス情報をもとに、上位コンピュータから該バックアップ対象ファイルシステムへのアクセスをしながら、該バックアップ対象ファイルシステムのデータのバックアップができるようにする。

【0016】さらに、本発明は、ディスクから読み出したデータを一時的に格納する半導体メモリを有するディスクドライブにおいて、該ディスクの記録領域の所定部分を書込み禁止領域に設定する第1の手段と、該半導体

メモリの記録領域を複数に分割して管理する第2の手段とを備え、該半導体メモリにおける該第2の手段によって分割形成された1乃至複数の記憶領域に、該第1の手段によって設定された書込み禁止領域の一部または全部から読み出したデータを記録して退避させ、しかる後に、該ディスクでの書込み禁止領域とした記録領域を書込み可能とする。

【0017】

【作用】コンピュータシステムにおけるアレイ上に配列された記憶装置には、ホストコンピュータが使用するデータ(ファイル)やこのデータから計算されるバリティ情報を格納するための情報記憶装置と、かかる情報記憶装置に障害が発生した場合にこの障害情報記憶装置と論理的に置き換えるための予備記憶装置とがある。本発明では、かかる予備記憶装置を、情報記憶装置でのデータをバックアップする際、情報記憶装置とデータバックアップ装置との間のデータ転送バッファとして用いる。

【0018】これによると、データバックアップ装置のデータ転送速度が情報記憶装置のデータ転送速度よりも充分遅くても、情報記憶装置からデータ転送バッファへのデータ転送速度が速いため、少なくとも情報記憶装置からデータ転送バッファへのデータ転送は高速で行なわれ、従って、情報記憶装置に関しては、データバックアップに要する時間が非常に短くなる。

【0019】現在大容量データバックアップ装置として一般に使用されている8ミリテープドライブやデータDAT(デジタル・オーディオ・テープレコーダ)装置のデータ転送速度は200kB/秒程度であり、ディスクドライブのデータ転送速度は3MB/秒程度であって、上記大容量データバックアップ装置と10倍以上の差がある。このため、予備ディスクドライブでバックアップ装置へのデータ転送をバッファリングすることにより、ディスクドライブでのデータバックアップのための占有時間を10分の1以下にすることができる。先に従来技術として挙げた例で説明すると、従来14時間程度かかっていたデータバックアップのためのディスクドライブの占有時間を1.4時間以下にすることができるということである。これは、1日に1回データバックアップを行なうシステムでは、このバックアップのためのシステム停止時間が1日当たり14時間から1.4時間に減少するということである。

【0020】また、本発明では、データバックアップ装置でバックアップされている情報記憶装置上でのバックアップ対象ファイルシステムへのアクセス情報を記録しておくことができるため、このバックアップ対象のファイルシステムを上位コンピュータ等によってアクセスされた後、このファイルシステムをデータバックアップ装置に転送することができる。従って、データバックアップ中でのアクセスがなされても、アクセス後のファイルシステムをバックアップすることができる。

【0021】さらに、本発明では、ディスクの記録領域でのデータバックアップしようとする領域を書込み禁止領域とし、この書込み禁止領域から読み出したデータを一時的に半導体メモリに格納して退避させる。この書込み禁止領域から全てのデータが読み出されて半導体メモリに退避されると、この領域の書込み禁止が解除されて通常の外部からのデータ書込みが可能となる。半導体メモリに退避されたデータはバックアップ装置に転送される。そこで、ディスクのデータが読み出されてから最終的にバックアップ装置に転送されるまでに要する時間であるデータバックアップ転送時間は充分長くても、ディスクから半導体メモリへのデータ転送は高速で行なわれるから、書込み禁止が設定されても短時間でそれが解除され、従って、データバックアップの占有時間が短い。

【0022】

【実施例】以下、本発明の一実施例を図面により説明する。図1は本発明によるディスクアレイ装置、データ記憶システム及びディスクアレイシステムのデータバックアップ方法の一実施例を示すブロック図であって、1はホストコンピュータ、1-1はCPU（中央処理ユニット）、1-2は主記憶部、1-3はI/Oバッファ、1-4はI/O（入出力）制御装置、1-5はディスクI/F（インターフェース）、1-6はOS（オペレーティングシステム）、1-7はファイル管理部、1-8はデバイスドライバ、2はディスクアレイコントローラ、2-1はホストI/F、2-2はCPU、2-3はデータ転送制御装置、2-4はパリティ生成/チェック部、2-5は転送バッファ、2-6-1、2-6-2、2-6-3、2-6-4は夫々ドライブI/F、3-1、3-2、3-3は夫々バックアップ装置、4-1、4-2は夫々スベアディスクドライブ、5はデータ/パリティディスクドライブ、6、7-1、7-2、7-3、7-4は夫々SCSI（Small Computer System Interface）バスである。

【0023】同図において、ホストコンピュータ1、ディスクアレイコントローラ2及びバックアップ装置3-1がSCSIバス6で相互に接続されており、ディスクアレイコントローラ2には、SCSIバス7-1、7-2、……、7-3により、ディスクドライブがアレイ状に接続されている。これらディスクドライブのうち、SCSIバス7-1、7-2、……夫々に接続されるディスクドライブはデータ/パリティディスクドライブ5であって、これらのうちのSCSIバス7-1、7-2、……の1つに接続されたディスクドライブが上記のパリティディスクドライブであって、残りがデータディスクドライブである。また、SCSIバス7-4に接続されているディスクドライブがスベアディスクドライブ4-1、4-2、……である。これらSCSIバス7-1、7-2、……、7-3夫々に接続されるディスクドライブの個数は等しく、各SCSIバス7-1、7-2、…

…の同順位のディスクドライブ（即ち、図面上で横方向に1列に並ぶデータ/パリティディスクドライブ5）が上記のパリティグループを構成していて、各パリティグループにSCSIバス7-3に接続されたスベアディスクドライブ4が1つずつ付加されている。

【0024】なお、バックアップ装置3-1の代りに、SCSIバス7-1～7-3のいずれか、もしくはディスクアレイコントローラ2に別個のSCSIバス7-4を設けてこれに、バックアップ装置3-3または3-2を設けるようにしてもよい。

【0025】以上のディスクアレイコントローラ2、アレイ状に配列されたディスクドライブ及びバックアップ装置によってディスクアレイ装置が構成されている。

【0026】ホストコンピュータ1は、演算などの処理を行なうCPU1-1、ディスクドライブ5などへの入出力データをバッファリングやキャッシングするI/Oバッファ1-3を有し、CPU1-1が使用するデータや命令を一時的に格納しておく主記憶部1-2、ディスクドライブ5などへのデータの入出力を制御するハードウェアであるI/O制御部1-4（最近のワークステーションでは、I/O制御部1-4に専用プロセッサを使用し、ディスクドライブなどへのデータの入出力処理のためのCPU1-1の負荷を軽減し、高速なI/O処理を実現するようにしている）、ディスクアレイコントローラ2とのインタフェース制御を行なうディスクI/F1-5及びOS1-6からなっている。データ/パリティディスクドライブ5などの2次記憶装置へのデータの入出力処理は、ソフトウェア的には、OS1-6のファイル管理1-7、デバイスドライバ1-8などのモジュールが管理する。これらはアプリケーションプログラムが要求するデータ/パリティディスクドライブ5とのデータ入出力のスケジューリング、バッファリング、アクセスデータ長などを決定する。

【0027】この実施例では、夫々のインターフェース（I/F）をSCSIとしている。SCSIはANSI（American National Standard Institute）で作成され、ISO（International Standard Organization）でも承認されたコンピュータ周辺機器用インタフェースであり、磁気ディスクドライブだけでなく、光ディスクドライブ、磁気テープ装置、プリンタ、スキャナなどもサポートしており、小型コンピュータに広く普及している。

【0028】ディスクアレイコントローラ2は、そのドライブI/F2-6-1、2-6-2、……、2-6-3に接続されてアレイ状に配置されたデータ/パリティディスクドライブ5やスベアディスクドライブ4を統括制御し、ホストコンピュータ1に恰も通常のディスクドライブが1台接続しているかのように見せかける。ホストコンピュータ1から送られるデータはホストI/F2-1から取り込まれ、その論理アドレスに従い、データ

転送制御部2-3、転送バッファ2-5を介してアレイ状に配置されたデータ/パリティディスクドライブ5に分配される。ホストコンピュータ1からデータの要求があると、データ転送制御部2-3、転送バッファ2-5を介してホストI/F2-1から、アレイ状に配置されたデータ/パリティディスクドライブ5に分配して配置されているデータが統合されてホストコンピュータ1に転送される。

【0029】また、ディスクアレイ装置では、同じパリティグループ内で1台のデータ/パリティディスクドライブ5が故障しても、そこでのデータの回復が行なえるようにするために、ホストコンピュータ1から送られてくるデータに冗長データ（即ち、パリティ情報）が付加され、データ/パリティディスクドライブ5のうちのパリティディスクドライブに格納する。このパリティ情報の生成とチェックはパリティ生成/チェック部2-4で行なわれる。データ/パリティディスクドライブが故障すると、そのパリティグループでのパリティディスクドライブに格納されているパリティ情報と故障していないデータディスクドライブのデータとで演算を行ない、故障したデータ/パリティディスクドライブのデータをこの計算によって生成して、このパリティグループに対するスペアディスクドライブ4に格納する。故障したデータディスクドライブでの全てのデータが回復してこのスペアディスクドライブ4に格納されると、このスペアディスクドライブは故障したデータディスクドライブの論理的位を占め、この故障したデータディスクドライブに代ってデータディスクドライブとなる。

【0030】このようにして、データディスクドライブの故障が発生しても、人手を介することなく、自動的に速やかにデータの回復処理を行なうことができる。なお、スペアディスクドライブは、データディスクドライブの故障が発生していない通常の場合、使用されない。

【0031】バックアップ装置3-1、3-2または3-3は、ディスクアレイ装置に格納されているデータをバックアップするためのものである。上記のように、ディスクアレイ装置では、1パリティグループ内で任意の1台のデータディスクドライブが故障してもデータを失うことがないが、同じパリティグループ内で複数台のデータディスクドライブが同時に故障する多重障害が発生した場合や、オペレータミスによって誤ってデータが消去された場合にはデータを失ってしまう。バックアップ装置3-1、3-2または3-3はこのようなことを防止するために設けられたものであって、これにディスクアレイ装置に格納されているのと同じデータをバックアップしておき、上記のような障害が生じたとき、バックアップ装置のデータをディスクアレイ装置に送って（バックアップデータの回復）使用できるようにする。かかるバックアップ装置としては、磁気テープ装置や光磁気ディスクドライブなど記憶容量当りのコストが安い可換

媒体の記憶装置が主に使用される。近年では、8ミリ磁気テープ装置で5GBなど大容量の装置が開発されている。

【0032】以下、この実施例の動作を説明する。

(1) ホストコンピュータ1がスペアディスクドライブとバックアップ装置とを認識する場合：図2はこの場合のホストコンピュータ1から見た場合の論理的な構成を示す図である。

【0033】同図において、ここでの符号2は、ディスクアレイが図1でのディスクアレイコントローラ2によって統括、制御され、ホストコンピュータ1からは1台のディスクドライブに見せかけられているデータ/パリティディスクドライブ5のアレイの総体を概念的に表わしたものであり、これをディスクアレイということにする。また、バッファディスク8は図1におけるスペアディスク4（一般には、複数台）を概念的に表わしたものである。ホストコンピュータ1からは、これらがバックアップ装置3（図1でのバックアップ装置3-1、3-2、3-3のいずれか）とともに1つのSCSIバスIDを持った装置として、図2に示すように見える。なお、SCSIバスIDとは、SCSIバス上に接続された複数台の装置を区別するための装置毎に割り振られたアドレスである。

【0034】以上の構成の場合、ホストコンピュータ1は、バッファディスク8を1つのSCSIバス装置として認識しているため、バックアップに必要な転送を全て自分の管理の元で行なうことができる。但し、必ずしも、全て管理する必要がないことはいうまでもない。

【0035】ディスクアレイ2のデータをバックアップ装置3にバックアップするのに行なうデータ転送は、ディスクアレイ2からバッファディスク8への転送(1)とバッファディスク8からバックアップ装置3への転送(2)とからなっている。好適な例としては、転送(1)はホストコンピュータ1がディスクアレイ2のファイル構造を認識しながら行ない、転送(2)は起動指示と終了確認のみホストコンピュータ1が介在する方法である。これによると、転送(2)中では、ホストコンピュータ1はディスクアレイ2に対して通常のアクセスを行なうことができる。

【0036】例えば、UNIXマシンの場合には、TARコマンドを用いてバッファディスク8への転送を行なう。この際、バッファディスク8はRAWデバイスとしてアクセスされる。TARコマンドで転送した場合、通常のMT装置へバックアップする場合と同じフォーマットになるため、データ転送(2)では、ホストコンピュータ1からバッファディスク8に対して転送すべき範囲と転送先さえ起動時に指示すればよい。但し、そのためには、バッファディスク8がSCSIバスコマンドの1つであるCOPYコマンドをサポートしていることが必要である。

【0037】上記転送(1)において、その開始から終了までホストコンピュータ1が介在する理由は、a) オープンしているファイルの処理、b) バックアップすべきデータが存在する領域の判断、c) ディスク上のデータフォーマットと通常のバックアップ装置(磁気テープ装置など。通常、これらはシーケンシャルアクセスデバイスであって、ランダムアクセスデバイスである磁気ディスクとは異なるフォーマットを持っている)のフォーマット変換などの処理を通常サポートしているコマンドのみで、上記転送(1)を行なうことができるためである。転送(1)をホストコンピュータ1の介在なしで行なう場合には、後に述べるように、上記a)、b)、c)の処理、判断ためのコマンド、メッセージなどをサポートする必要がある。

【0038】但し、図1に示すように、ディスクアレイコントローラ2は、実際には1つのSCSIバス装置であるのに、ホストコンピュータ1に対しては、図2におけるディスクアレイ2とバッファディスク8の二役をしなければならない。SCSIバス規約にはこのような構成は存在しないが、実施は可能である。

【0039】(2)ホストコンピュータ1は、バックアップ装置を認識するが、スベアディスク装置を認識しない場合：図3はこの場合のホストコンピュータから見た論理的構成図であり、図2に対応する部分には同一符号を付けている。

【0040】ここでは、図2でのバッファディスク8はホストコンピュータ1によって認識されていない。この場合、上記のデータ転送(1)に当たるデータ/パリティディスクドライブ5からスベアディスクドライブ4へのデータ転送は、ディスクアレイコントローラ2が起動、終了処理を含めて全て管理する。この場合、通常のバックアップ用のコマンド(上記の例では、TARコマンド)を用いてバックアップを行なうことも考えられる。それでも、一旦スベアディスクドライブ4でバッファリングすることにより、通常のアクセスのためにデータ/パリティディスクドライブ5を使用できる時間が増大する効果がある。

【0041】また、バックアップのために必要なファイル管理情報を読み込むためには、ディスクアレイ2でのデータをホストコンピュータ1が読み込むが、バックアップ装置3にデータを転送するためには、ホストコンピュータ1にデータを読み込まず、SCSIバスのCOPYコマンドを用いてディスクアレイ2のデータをバックアップする方法がある。COPYコマンドを受け取ったときには、スベアディスクドライブ4をデータ転送バッファとして用い、ディスクアレイコントローラ2が制御を行なうことにより、より高速なバックアップを行なうことができる。

【0042】ディスクアレイコントローラ2でのバックアップ処理は、バックアップ装置3の物理的接続位置に

より、図4、図5及び図6に示す3種類の好適な例を挙げることができる。

【0043】図4はバックアップ装置3をホストI/F2-1側のSCSIバス6に接続した場合であり、図5は他のディスクドライブが接続されていないドライブI/F2-6に接続した場合であり、図6はスベアディスクドライブ4が接続されているドライブI/F2-6に接続した場合である。なお、図4～図6においても、データ/パリティディスクドライブ5及びスベアディスクドライブ4が夫々1つずつ示されているが、勿論これらは図1で説明したように複数台アレイ状に配置されており、ディスクアレイコントローラ2によって1台のディスクドライブのように統括制御されているものを概念的に表わしているものである。

【0044】図4に示す例の場合、ホストコンピュータ1はバックアップを行なうデータ/パリティディスクドライブ5の領域のファイル管理情報を読み込み、データ/パリティディスクドライブ5からバックアップ装置3への転送を指示するコマンドをディスクアレイコントローラ2に対して発行する(その転送はホストコンピュータ1が介在してもよい)。ディスクアレイコントローラ2は転送先がバックアップ装置3であるデータを一旦スベアディスクドライブ4に転送し、次に、このスベアディスクドライブ4からバックアップ装置3への転送を行なう。これにより、データ/パリティディスクドライブ5がバックアップのために占有される時間を減らすことができる。

【0045】図5に示す例の場合には、バックアップ装置3がディスクアレイコントローラ2のドライブI/F2-6に接続されているため、スベアディスクドライブ4からバックアップ装置3へのデータ転送はホストI/F2-1側のSCSIバス6を通らずに行なわれる。このため、バックアップ中のホストコンピュータ1からデータ/パリティディスクドライブ5へのアクセス性能の劣化が小さい。

【0046】但し、この場合、ホストコンピュータ1に対して図3に示したようにバックアップ装置3が接続されていると見せるためには、ディスクアレイコントローラ2は、バックアップ装置3に対するホストコンピュータ1からのアクセスを透過的に受渡してやらなければならない。このことは、図2におけるバッファディスク8と同様、SCSIバスの規約には規定されていない方法ではあるが、実施は可能である。

【0047】図6に示す例は、バックアップ装置3をスベアディスクドライブ4と同じSCSIバス上に接続する場合である。この場合には、スベアディスクドライブ4からバックアップ装置3へのデータ転送がディスクアレイコントローラ2の内部も通過しないで行なわれ、このために、さらにバックアップ中のホストコンピュータ1からデータ/パリティディスクドライブ5へのアクセ

ス性能劣化を防止できる。かかる効果は、スペアディスクドライブ4がCOPYコマンドをサポートすることによって最大になる。

【0048】(3) ホストコンピュータ1がバックアップ装置もスペアディスクも認識しない場合：図7はこの場合のホストコンピュータ1からみた概念的な構成図である。実際の物理的な接続は、バックアップ装置3を認識する場合と同様に図4、図5及び図6に示す3種類が好適な実施例として挙げられる。この場合、ホストコンピュータ1がバックアップ装置3を認識しないため、ディスクアレイコントローラ2はバックアップ用のコマンドをサポートしなければならない。なお、かかるコマンドはSCSIバスで規定されていないが、このようなコマンドでも、ベンダユニークとして実施することができる。バックアップを行なうためには、まず、ホストコンピュータ1がバックアップ対象領域のファイル管理情報を読み込み、バックアップすべきアドレスとバックアップを指示するコマンドをバックアップを行なう順にディスクアレイ2に発行してやればよい。

【0049】(4) 2ポートSCSIバスディスクドライブを使用した場合：以上説明した各実施例では、ディスクドライブ装置として通常の1ポートのものを想定していた。しかし、以上全ての実施例において、2ポートSCSIバスディスクドライブを使用した構成も考えられる。その一実施例を図8に示す。かかる実施例の特徴は、データ/パリティディスクドライブ5からスペアディスクドライブ4へのデータ転送やコマンド/メッセージの転送が、通常のホストコンピュータ1からデータ/パリティディスクドライブ5へのアクセスとSCSIバス上で競合しない点である。スペアディスクドライブ4からバックアップ装置へのデータ転送も、バックアップ装置3が点線で示されている位置に接続されている場合には、ホストコンピュータ1からデータ/パリティディスクドライブ5へのアクセスと競合することがなく、バックアップ中の性能劣化が少ない。

【0050】(5) システムを閉塞せずにバックアップを行なう方法：以上説明した実施例では、バックアップのためにディスクドライブを占有する時間を短縮することができた。バックアップを行なっているファイルシステムもしくはパーティションは、バックアップ中データ内部の統一性や管理情報との統一性が失われないようにするために、少なくとも一纏まりのファイルとその管理情報の処理中、その領域へのライトアクセスを停止（システムの閉塞）しなければならない。しかし、オンラインランザクションの用途では24時間のサービスが要求される場合もあり、このようなシステムでは、上記のような閉塞は許容できない。

【0051】このようなシステムの閉塞を行わずにバックアップを行なうためには、以下の手順に従えばよい。即ち、1) バックアップ中にバックアップ対象のデ

ィスク領域の中でオープンされているファイルを記録する。記録先はディスク上でも、メモリ上でもよい。オープンしているファイルのリストは、一般に、ホストコンピュータの主記憶部に存在する。2) 対象領域をオープンされているファイルも含めて（勿論、省いてもよい）、バックアップ装置にバックアップする。3) ファイルがクローズされたときバックアップ対象領域であったファイルならば、差分バックアップの要領で、2)で行なったバックアップに連続した領域にバックアップを行なう。

【0052】なお、1)の代りに、ファイルオープン処理以降の全てのライトアクセスを記録し、ライトアクセスのない領域は、オープンされていても、バックアップを行ない、ライトアクセスのあった領域のみ、クローズ処理後、差分バックアップを行なうようにしてもよい。

【0053】図9は本発明によるディスクドライブの一実施例を示すブロック図であって、8a、8bはSCSIバス、9a、9bはSCSI制御部、10はメモリ制御部、11はデータメモリ、12はフォーマット制御部、13はリード/ライト用アンプ（以下、R/Wアンプという）、14は磁気ヘッド、15は記録ディスク、16は書込禁止領域判定部、17はスピンドルモータ、18はVCM、19はバッテリーである。

【0054】図9において、SCSI制御部9aはSCSIバス8aのプロトコル制御を行ない、SCSI制御部9bはSCSIバス8bのプロトコル制御を行なう。メモリ制御部10は、SCSI制御部9a、9bとデータメモリ11とフォーマット制御部12との間のデータ転送を制御し、図10に示すように、データメモリ11を4つの記憶領域に分割して管理する。データメモリ11は少なくともディスク15の数トラック分の容量を有していることが望ましく、ここでは、その容量は1Mバイトとする。データメモリ11は、バッテリー19により、データの揮発化を行なっている。フォーマット制御部12は、ディスク15上の記録トラックのセクタ分割方法を管理する。R/Wアンプ13は、ディスク15へのデータ記録時、データによって磁気ヘッド14を駆動し、ディスク15からのデータ再生時、磁気ヘッド14からの再生データ信号を増幅してデジタル信号に変換する。書込禁止領域判定部16は、ディスク15上のデータ書込み予定領域が書込み禁止領域か否かを判定し、書込み禁止領域ならば、フォーマット制御部12の書込み動作を停止させる。

【0055】次に、ディスク15上の記録領域の所定領域のデータをバックアップする場合のこの実施例の動作を図11を用いて説明する。

【0056】ここで、データバックアップ動作は、バックアップの領域をデータメモリ11へ退避させるためのロック命令と、データメモリ11に退避したデータを外部のバックアップ装置に転送させるためのバックアップ

リード命令との2種類の命令をディスクドライブが図示しないホスト装置から受け取るにより、実行される。

【0057】ディスクドライブは、ホスト装置から部分ロック命令を受けると(ステップ101)、ディスク15のこの部分ロック命令で指定されるロック領域を一時的に書き込み禁止領域に設定し(ステップ102)、かつこのロック領域中のデータを一時的に退避する領域をデータメモリ11内に確保する(103)。データメモリ11は、通常、図10の「通常時」として示す状態にあるが、部分ロック命令があると、ロック領域の容量がデータメモリ11の容量の3/4以下ならば、ロック領域のデータを格納できる最小領域を退避領域として、図10に示す「退避時1」、「退避時2」、「退避時3」のいずれかの状態に設定される。ロック領域の容量がデータメモリ11の容量の3/4を越えるならば、データメモリ11を図10の「退避時3」の状態とし、データメモリ11の容量の3/4に等しい容量のデータをロック領域から読み出し、データメモリ11の第2～第4領域の退避領域に格納し(ステップ104)、読出しが終了したディスク15のロック領域の書き込みを許可する(ステップ105)。データメモリ11の退避領域にロック領域の全データを格納できない場合には(ステップ106)、部分ロックの完了をSCSI制御部10に通知する(ステップ107)。また、データメモリ11の退避領域にロック領域の全データを格納できた場合には(ステップ106)、ロック対象領域の退避が完了すると、書き込み可能であることを通知する(ステップ108)。

【0058】その後、バックアップリード命令を受けると(ステップ109)、ディスクドライブは、このバックアップリード命令が転送されてきたSCSIバスがSCSIバス8a、8bのいずれかを介し、データメモリ11の退避領域からバックアップデータをホスト装置へ転送する(ステップ110)。ホスト装置からバックアップデータの受信完了の通知を受けると(ステップ111)、ディスクドライブは、前記の部分ロック命令で指定されたバックアップ領域のバックアップデータの全てをホスト装置へ転送終了したならば(ステップ112)、データメモリ11に設定された退避用領域を通常用領域に戻し(ステップ113)、ディスク15での部分ロック状態を解除し(ステップ114)、バックアップ動作を完了する。

【0059】バックアップ領域に未転送部分が残っているときには(ステップ112)、引き続きステップ104～112の一連の動作を行ない、残りのロック領域のデータに対してデータメモリ11への退避(ステップ104)とホスト装置への転送(ステップ110)を繰り返す。

【0060】なお、ステップ107、108を経た後、バックアップリード命令がないときに(ステップ10

9)、ディスク15への読出し命令を受けると(ステップ120)、通常読出し処理が行なわれ、ディスク15への書き込み命令を受けると(ステップ122)、書き込み禁止領域であるときには(ステップ124)、書き込みを禁止するが(ステップ126)、書き込み禁止領域でないときには通常書き込み処理を行なう(ステップ125)。読出し、書き込み命令のいずれをも受けないときには、他の命令の処理を行なう(ステップ123)。

【0061】以上のようにして、この実施例では、バックアップデータをデータメモリ11に退避した後は、ディスク15へのデータの書き込み動作とバックアップのためのディスクドライブのスループットの低下を抑えることができる。

【0062】図12は図9～図11で説明したディスクドライブをアレイ上に配列してなる本発明によるディスクアレイ装置のさらに他の実施例を示すブロック図であって、20はホストコンピュータ、21はディスクアレイ装置、22はホストインタフェース制御部(以下、ホストI/Fという)、23はデータ分配制御部、24はキャッシュメモリとしてのバッファメモリ、25はパリティ生成部、26はデータ回復部、27は磁気テープ装置制御部(以下、MT制御部という)、28は磁気テープ装置(以下、MT装置という)、29a～29eはSCSI制御部、30a～31e、32は図9で説明したディスクドライブ(以下、HDDという)、33、34はSCSI制御部である。

【0063】図12において、ホストI/F22は、ホストコンピュータ20からディスクアレイ装置21への命令や書き込みデータの受信と、ディスクアレイ装置21からホストコンピュータ20への命令実行結果や読出しデータの送信を行なう。データ分配制御部23は、ホストI/F22とキャッシュメモリ24とパリティ生成部25とデータ回復部26とMT制御部27とSCSI制御部29a～29e、33、34との間のデータ転送の制御を行なう。さらに、このデータ分配制御部23は、データ書き込み時には、ホストコンピュータ20からの書き込みデータを4kバイト単位のデータブロック(先のデータストライプ)に分割して、HDD31a～31eへ振り分け、逆に、データ読出し時には、HDD30a～31eからのデータブロックをまとめて、ホストコンピュータ20への読出しデータに再構成する。

【0064】各データブロックは次のように書き込まれる。即ち、いま、データが4個のデータブロックD00、D01、D02、D03に分割されたとすると、図13に示すように、データブロックD00がHDD30aに、データブロックD01がHDD30bに、データブロックD02がHDD30cに、データブロックD03がHDD30dに順に配置されることになり、HDD30eには、これらデータブロックD00、D01、D02、D03から生成したパリティブロックP0が配置さ

れる。次のデータがデータブロックD10, D11, D12, D13に分割され、それらのパリティブロックをP1とすると、これらも同様にHDD30a~30eに記憶されるが、パリティブロックP1はHDD30dに記憶される。以下同様にして、順次のデータが記憶される。なお、同一データから分割されたデータブロックとこれらによって生成されたパリティブロックとをまとめてパリティグループということは、先に述べたとおりである。

【0065】パリティブロックは、パリティ生成部13が、次の数2に示すように、同一パリティグループの4つのデータブロックを排他論理することにより、生成される。

【0066】

【数2】

【数2】

$$P0 = D00 \oplus D01 \oplus D02 \oplus D03$$

【0067】キャッシュメモリ12は書き込みデータや読出しデータを一時的に格納するものである。また、データ回復部14は、HDDの故障のためにパリティグループ内のデータブロックが損失した場合、パリティグループ内の正常なデータブロックとパリティブロックを用いて、損失したデータブロックを回復するためのものである。例えば、HDD30aが故障した場合、次の数3に示すように、データブロックD01, D02, D03とパリティブロックP0とを用いて、データブロックD00を回復できる。

【0068】

【数3】

【数3】

$$D00 = D01 \oplus D02 \oplus D03 \oplus P0$$

【0069】SCSI制御部29a~29e, 33, 34はHDDを接続しているSCSIバスのプロトコル制御を行なう。MT制御部27はバックアップ装置としてのMT装置28を制御し、HDDからのデータをバックアップのため、MT装置16内の磁気テープに記録する。HDD30a~31e, 32は2系統のSCSIバスに接続可能である。また、HDD33はスペア用である。

【0070】以下、この実施例のデータ記録再生動作を図14を用いて説明するが、まず、ホストコンピュータ20からのデータをHDD30aのD00領域とHDD30bのD01領域とに記録する場合を説明する。

【0071】時刻t01にホストコンピュータ20から書き込み命令と書き込みデータが送信されると、ホストI/F22はこれら書き込み命令と書き込みデータをデータ分配制御部23へ転送する。時刻t02で、データ分配制御部23は書き込みデータをキャッシュメモリ24へ転送

し、また、パリティブロックの生成のために、SCSI制御部29a, 29b及び29eへデータ読出しを要求する。そこで、SCSI制御部29aはHDD30aへデータブロックD00の読取りを要求し、SCSI制御部29bはHDD30bへデータブロックD01の読出しを要求し、SCSI制御部29eはHDD30eへパリティブロックP0の読出しを要求する。

【0072】時刻t03で、HDD30a, 30b, 30eは、夫々、同一パリティグループのデータブロック、パリティブロックを読み取り、SCSI制御部29a, 29b, 29eへこれらデータブロックを転送する。そこで、SCSI制御部29a, 29b, 29eは、これら読取りデータブロックをデータ分配制御部23へ転送し、データ分配制御部23はこれらをキャッシュメモリ12へ格納する。

【0073】ここで、HDD30a, 30b, 30eからの読取りデータブロックとパリティブロックを、夫々、rd0, rd1, rp0とし、ホストコンピュータ20からの書き込みデータのデータブロックをwd0, wd1とする。

【0074】時刻t04で、キャッシュメモリ24から、データブロックrd0, rd1, パリティブロックrp0, データブロックwd0, wd1がパリティ生成部25へ転送される。パリティ生成部25では、次の数4のように、これらから新しいパリティブロックwp0が生成され、キャッシュメモリ24へ転送される。

【0075】

【数4】

【数4】

$$wp0 = rd0 \oplus rd1 \oplus rp0 \oplus wd0 \oplus wd1$$

【0076】時刻t05で、データ分配制御部23は、SCSI制御部29a, 29b, 29eへ、書き込み命令と、夫々、書き込みデータブロックwd0, wd1とパリティブロックwp0を送る。これにより、SCSI制御部29aはHDD30aへデータブロックwd0の記録領域D00への書き込みを要求し、SCSI制御部29bはHDD30bへデータブロックwd1の記録領域D01への書き込みを要求し、SCSI制御部29eはHDD30eへパリティブロックwp0の記録領域P0への書き込みを要求する。

【0077】時刻t06で、HDD30a, 30b, 30eのライト動作の完了がSCSI制御部29a, 29b, 29eを通じてデータ分配制御部23へ通知され、これにより、データ分配制御部23は、ホストI/F22を経由して、ホストコンピュータ20へ書き込み命令の完了を通知する。以上により、データ記録動作は終了する。

【0078】次に、例えば、SCSI制御部29bまたはそれに接続されるSCSIバスが故障した場合の、H

DD30aのD00領域とHDD30bのD01領域とからのデータの再生動作を説明する。

【0079】いま、時刻t10でかかる故障が生じたとすると、SCSI制御部29bはこのことをデータ分配制御部23へ通知する。その後、時刻t11で、ホストコンピュータ20から読取り命令が送信されるたとして、ホストI/F22は読取り命令をデータ分配制御部23へ転送する。時刻t12で、データ分配制御部23はSCSI制御部29aへHDD30aの読取りを要求する。さらに、データ分配制御部23は、故障中のSCSI制御部29bの代わりに、SCSI制御部33へHDD30bの読取りを要求する。

【0080】時刻t13で、HDD30a、30bが、夫々、SCSI制御部29a、33を介して、データ分配制御部23へ読取りデータブロックを転送する。そこで、データ分配制御部23はHDD30a、30bからのデータブロックを供給し、ホストコンピュータ20へ送信する。これにより、データ再生動作が終了する。

【0081】次に、この実施例のデータ回復動作を図15を用いて説明する。時刻t20でHDD30aが故障したとすると、HDD30aはこれを検出し、SCSI制御部29aを介して、データ分配制御部23へこの旨通知する。時刻t21で、データ分配制御部23は、データ回復部26へHDD30a内のデータブロックのスペアHDD32への回復を要求する。データ回復部26は、時刻t22で、SCSI制御部33を介し、HDD30b~30eへ、先頭のパリティグループのデータブロックD01、D02、D03及びパリティブロックP0の読取りを要求する。

【0082】時刻t23から、データ回復部26は、HDD30a~30eからの読取りデータを受け、順次上記数2の演算を実行し、データブロックD00を回復させる。そして、時刻t24で、先頭のパリティグループの各ブロックD01、D02、D03、P0の転送とデータ回復演算が完了すると、データ回復部26は、SCSI制御部33を介し、スペアHDD32へ回復したデータブロックD00を書き込む。時刻t24で、スペアHDD32での書き込みが完了すると、この旨をSCSI制御部33を介してデータ回復部26へ通知する。データ回復部26は、データ分配制御部23へデータブロックD00の回復完了を通知し、データブロックD00のアクセスを許可する。引き続き他のデータブロックについて、順次回復処理を行なう。

【0083】時刻t26で、データ回復部26は、最後のデータブロックDk0の回復演算を完了し、SCSI制御部33を介してスペアHDD32へ回復したデータブロックDk0を書き込む。そして、時刻t27で、スペアHDD32が書き込み完了をSCSI制御部33を介してデータ回復部26へ通知する。データ回復部26は、データ分配制御部23へ全データブロックが回復完

了したことを通知し、データ回復処理が完了する。

【0084】以上のデータ回復処理においては、SCSI制御部29a~29eが使用されないため、データ回復処理中にも、故障したHDD30aのパリティグループ以外のHDD31a~31eには並行してデータブロックの書き込み/読取りが可能であり、HDD30b~30eには読取りが可能である。

【0085】次に、この実施例のデータバックアップ動作を図16を用いて説明する。時刻t30で、ホストコンピュータ20からディスクアレイ装置21内のデータブロックD12~D53の領域の部分バックアップ命令があったとする。そこで、データ分配制御部23は、SCSI制御部29a~29eを介して、HDD30a~30eへデータブロックD12~D53の部分ロックを要求し、ホストコンピュータ20からのかかる領域への書き込み要求の受付を禁止する。各HDDは、前記のように1Mバイトのデータメモリ11(図9)を持つため、このデータメモリ11の退避領域にロック領域のデータを全て格納して退避させることができ、このロック対象領域の各ブロックの退避が完了すると、かかる領域での書き込みが可能であることを通知する。この実施例では、1つのパリティグループで最大3840kバイトのデータを退避できる。

【0086】次に、時刻t32で、SCSI制御部20a~20eの全てがHDDの部分ロックの完了を、データ分配制御部23へ通知し、これにより、データ分配制御部23はSCSI制御部33にHDD30a~30eからMT装置28へのバックアップデータの転送を要求する。さらに、ホストコンピュータ20から指定されたバックアップ領域が全てHDD30a~30e内のデータメモリ11内に格納できた場合には、ホストコンピュータ20からのこの領域への書き込み要求の受付を許可する。

【0087】SCSI制御部33は、時刻t33に、部分バックアップの先頭データブロックD12の読取りをHDD30cへ要求する。データブロックD12を受けると、SCSI制御部33は、時刻t34に、MT制御部27へデータブロックD12を送り、MT装置28へ書き込みを要求する。これにより、MT制御部27はデータブロックD12のMT装置28への書き込みを開始する。

【0088】SCSI制御部33は、データブロックD12の転送が終了すると、時刻t35に、HDD30dへデータブロックD13の読取りを要求し、データブロックD12のMT装置28への書き込み完了の前に、予め、次のデータブロックD13をSCSI制御部33へ読み出しておく。時刻t36に、MT制御部27からデータブロックD12の書き込み完了通知があると、SCSI制御部33はHDD30cへデータブロックD12の受信完了を通知する。

【0089】続いて、SCSI制御部33は、読み取った上記のデータブロックD13のMT装置28への書込み要求と、データブロックD20のHDD30aからの読取り要求とを出す。時刻t37に、MT制御部27からデータブロックD13の書込み完了通知を受けると、SCSI制御部33は、HDD30dへデータブロックD13の受信完了を通知し、データ分配制御部23に第1列のバリティグループのデータブロックのバックアップ終了を通知する。

【0090】同様に、時刻t38に、MT制御部27からデータブロックD23のライト完了通知を受けると、SCSI制御部33は、データ分配制御部23に第2列のバリティグループのデータブロックのバックアップ終了を通知する。時刻t39に、MT制御部27からバックアップ領域の末尾のデータブロックD53のライト完了通知を受けると、SCSI制御部33は、データ分配制御部23に第2列のバリティグループのデータブロックのバックアップ終了を通知し、時刻t40に、ホストコンピュータ20に部分バックアップの完了を通知する。

【0091】以上、バックアップ領域がHDD内のデータメモリ11（図9）に全て退避できる場合を説明したが、バックアップ領域の方が大きい場合には、データ分配制御部23は、ディスク15（図9）のバックアップ領域のうち、データメモリ11に退避された領域からホストコンピュータ20からの書込み要求の受付を許可する。

【0092】以上のように、バックアップデータの転送に際しては、SCSI制御部29a～29eは使用されないため、バックアップ処理中にも、並行して、HDD30a～31eへのアクセスが可能である。

【0093】この実施例では、通常のHDDアクセス用バスに加えて、データ回復及びデータバックアップ用に使用するバスを持っているため、データ回復またはデータバックアップのためのバスの占有による性能低下を防ぐことができる。さらに、データ回復及びデータバックアップ用に使用するバスはスベア用HDDにも使用するため、スベアHDD32の制御のために別にSCSI制御部を設ける必要がなく、この分コストを削減できる。

【0094】また、この実施例では、1つのバリティグループを1本のバスで接続しているため、HDDを順次読み出すだけで、バリティグループ内の複数のHDDに分散して配置されたデータブロックを元の順序に並べることができる。

【0095】

【発明の効果】以上説明したように、本発明によれば、バックアップ中のデータディスクの占有時間をバックアップ装置と磁気ディスクドライブの転送速度の比だけ減らすことができる（典型的な例では、磁気テープ装置が約200kB/secであるのに対し、磁気ディスクド

ライブでは3MB/secと1/15程度である）。

【0096】また、本発明によれば、バックアップ対象領域のオープンされているファイルを記録し、クローズ後に再度バックアップを行なうことにより、システムを閉塞せずにバックアップを行なうことができる。

【0097】さらに、本発明によれば、ディスクドライブ内の半導体メモリの一部にバックアップデータを退避しておき、通常のディスクアクセスの合間に、バックアップ用の装置に転送するので、データバックアップ時にディスクドライブが長時間占有されることを防ぐことができる。また、ディスクのバックアップ領域のうちデータ退避の終わっていない領域は書込み禁止となっているため、そこに書き込まれているデータを誤って壊すこともない。

【図面の簡単な説明】

【図1】本発明によるディスクアレイ装置、データ記憶システム及びディスクアレイシステムのデータバックアップ方法の一実施例を示すブロック図である。

【図2】図1に示した実施例でのホストコンピュータ側からみた論理的接続の一具体例を示す構成図である。

【図3】図1に示した実施例でのホストコンピュータ側からみた論理的接続の他の具体例を示す構成図である。

【図4】図3に示した論理的接続でのディスクアレイコントローラ側の処理の一具体例を示す図である。

【図5】図3に示した論理的接続でのディスクアレイコントローラ側の処理の他の具体例を示す図である。

【図6】図3に示した論理的接続でのディスクアレイコントローラ側の処理のさらに他の具体例を示す図である。

【図7】図1に示した実施例でのホストコンピュータ側からみた論理的接続のさらに他の具体例を示す構成図である。

【図8】本発明によるディスクアレイ装置、データ記憶システム及びディスクアレイシステムのデータバックアップ方法の他の実施例を示すブロック図である。

【図9】本発明によるディスクドライブの一実施例を示すブロック図である。

【図10】図9でのデータメモリのデータバックアップ動作中の状態を示す図である。

【図11】図9に示した実施例のデータバックアップ時の動作を示すフローチャートである。

【図12】図9に示したディスクドライブを用いた本発明によるディスクアレイ装置のさらに他の実施例を示すブロック図である。

【図13】図12における各ディスクドライブでのデータ書込み状態を示す図である。

【図14】図12に示した実施例のデータ書込み、データ読出し動作を示すタイミング図である。

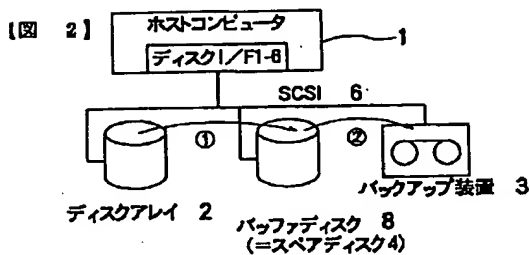
【図15】図12に示した実施例でのデータ回復動作を示すタイミング図である。

【図16】図12に示した実施例でのデータバックアップ時の動作を示すタイミング図である。

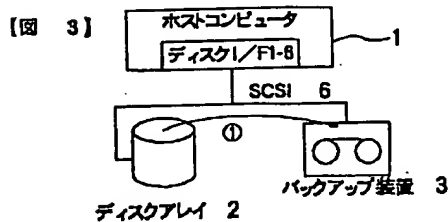
【符号の説明】

- | | | | |
|----|------------------|------------------|-----------|
| 1 | ホストコンピュータ | 15 | ディスク |
| 2 | ディスクアレイコントローラ | 16 | 書込禁止領域判定部 |
| 3 | バックアップ装置 | 20 | ホストコンピュータ |
| 4 | スペアディスクドライブ | 21 | ディスクアレイ装置 |
| 5 | データ/パリティディスクドライブ | 23 | データ分配制御部 |
| 6 | ホスト側SCSIバス | 24 | バッファメモリ |
| 7 | ドライブ側SCSIバス | 25 | パリティ生成部 |
| 10 | メモリ制御部 | 26 | データ回復部 |
| 11 | データメモリ | 28 | 磁気テープ装置 |
| 12 | フォーマット制御部 | 29a~29e | SCSI制御部 |
| | | 30a~30e, 31a~31e | ディスクドライブ |
| | | 33, 34 | SCSI制御部 |

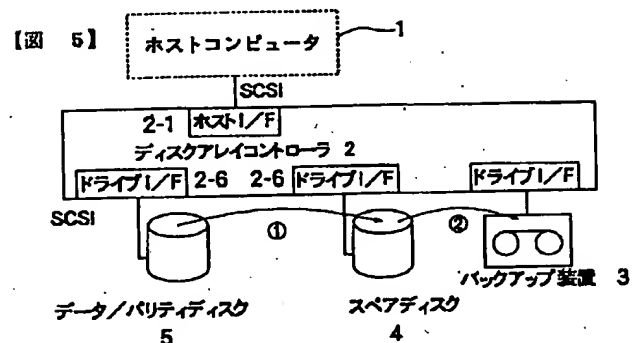
【図2】



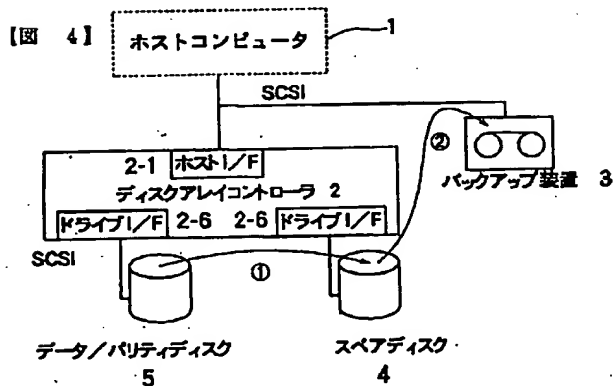
【図3】



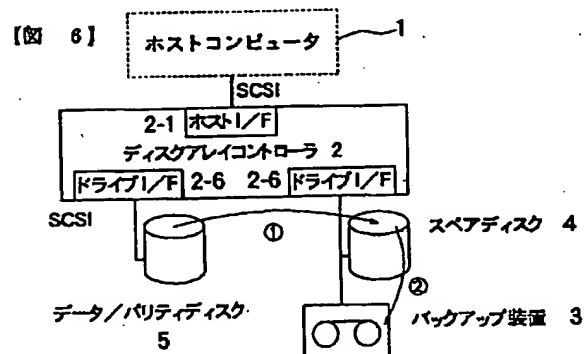
【図5】



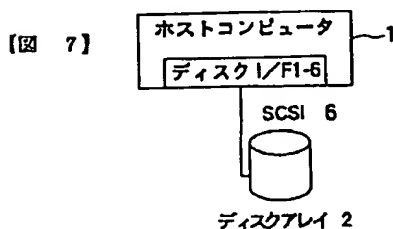
【図4】



【図6】

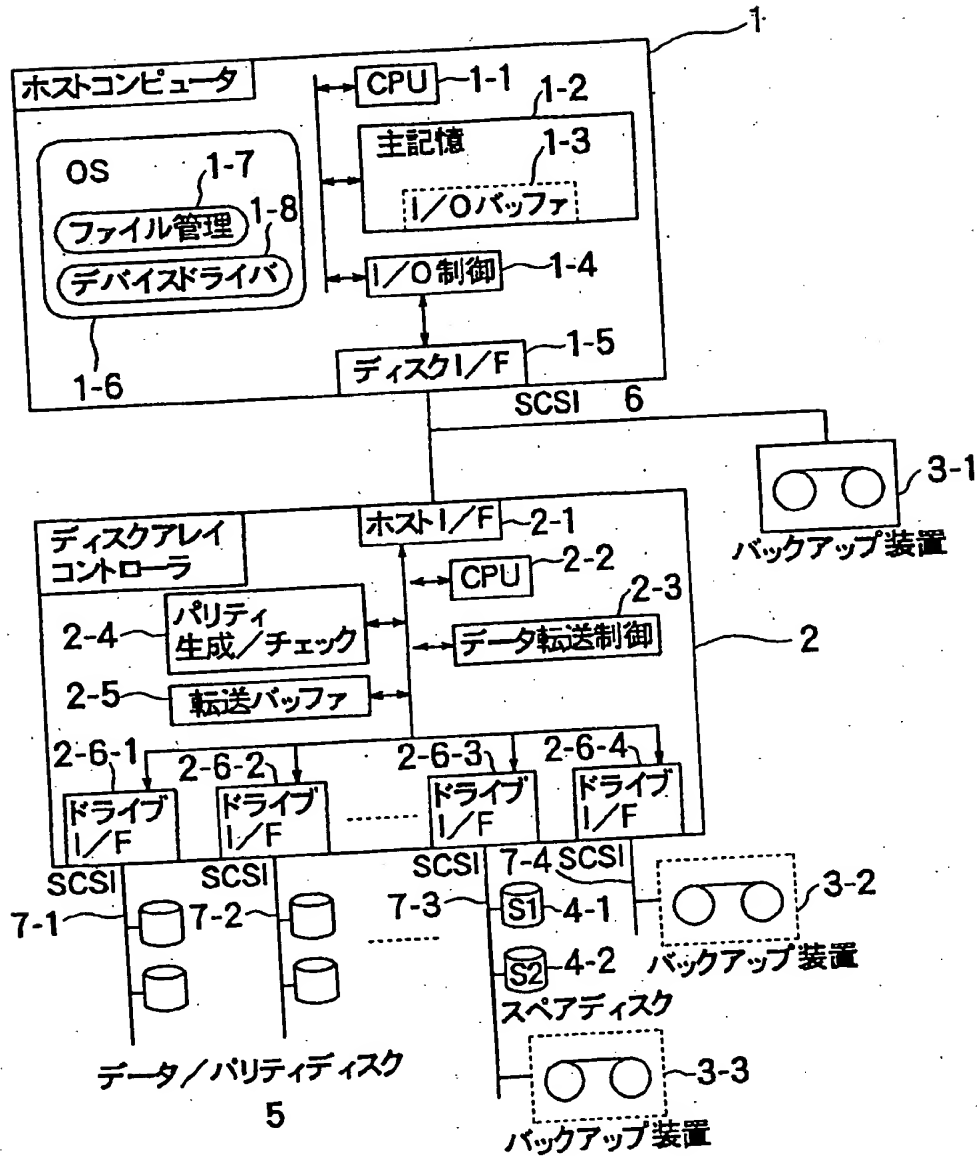


【図7】



【図1】

【図 1】



【図13】

【図10】

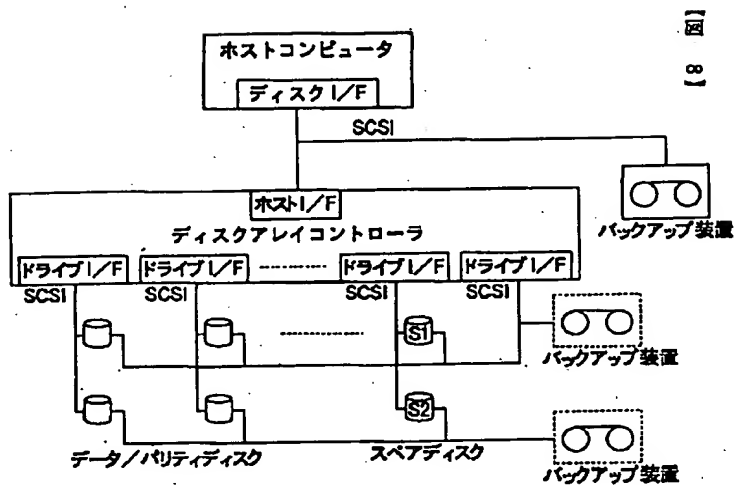
【図10】

第1領域	通常用	通常用	通常用	通常用
第2領域	通常用	通常用	通常用	退避用
第3領域	通常用	通常用	退避用	退避用
第4領域	通常用	退避用	退避用	退避用
	通常時	退避時1	退避時2	退避時3

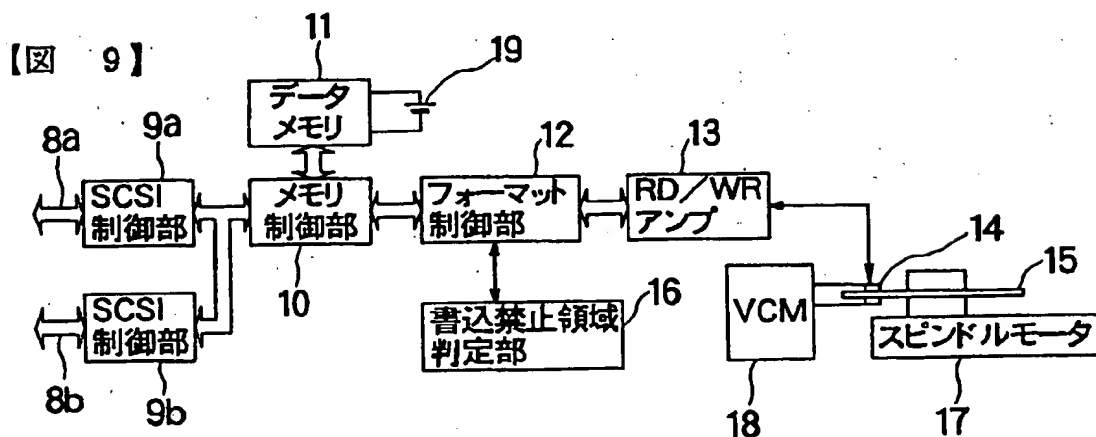
【図13】

HDD30a	HDD30b	HDD30c	HDD30d	HDD30e
D00	D01	D02	D03	P0
D10	D11	D12	P1	D13
D20	D21	P2	D22	D23
D30	P3	D31	D32	D33
P4	D40	D41	D42	D43
D50	D51	D52	D53	P5
Dk0	Dk1	Dk2	Pk	Dk3

【图8】

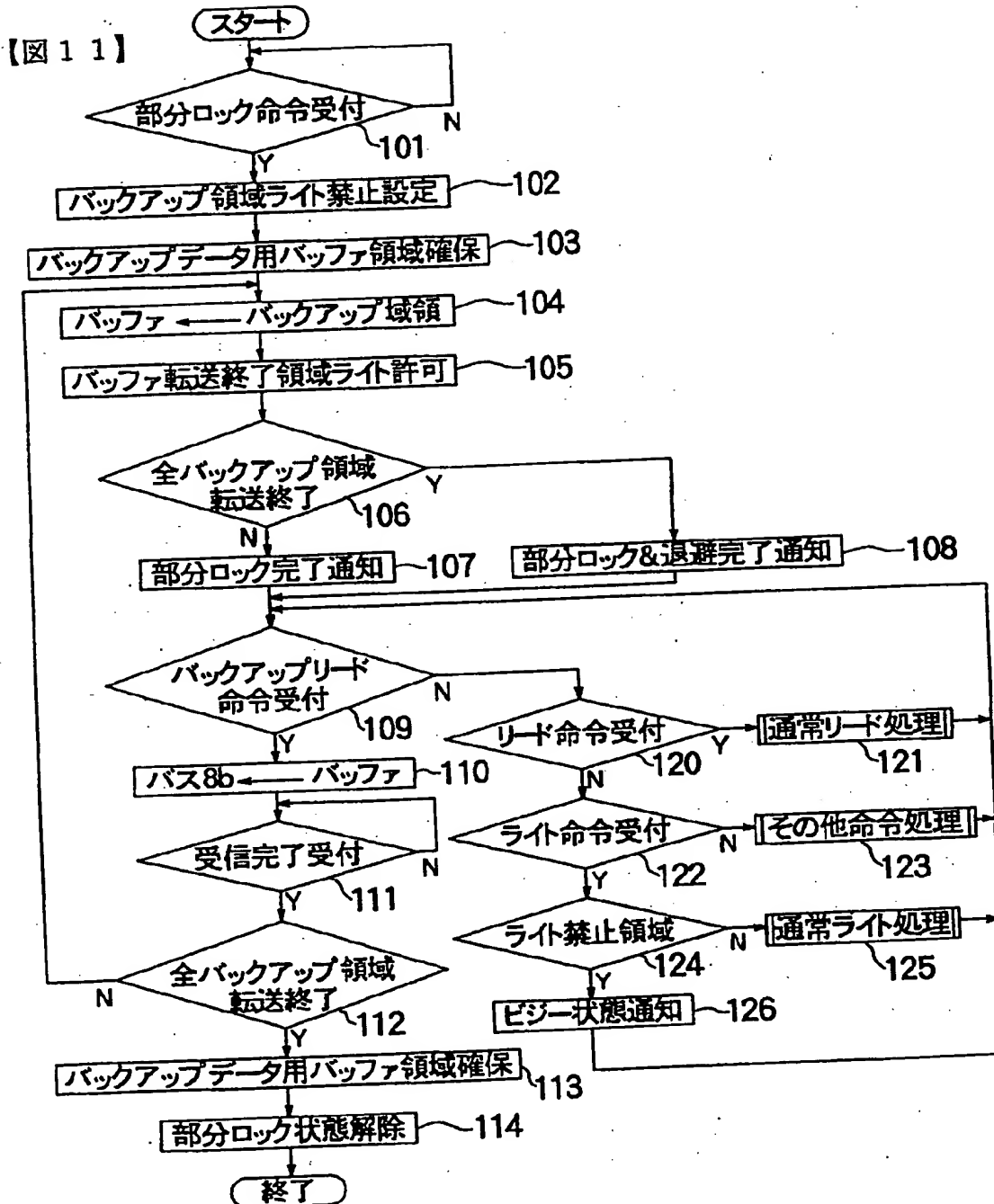


【図9】

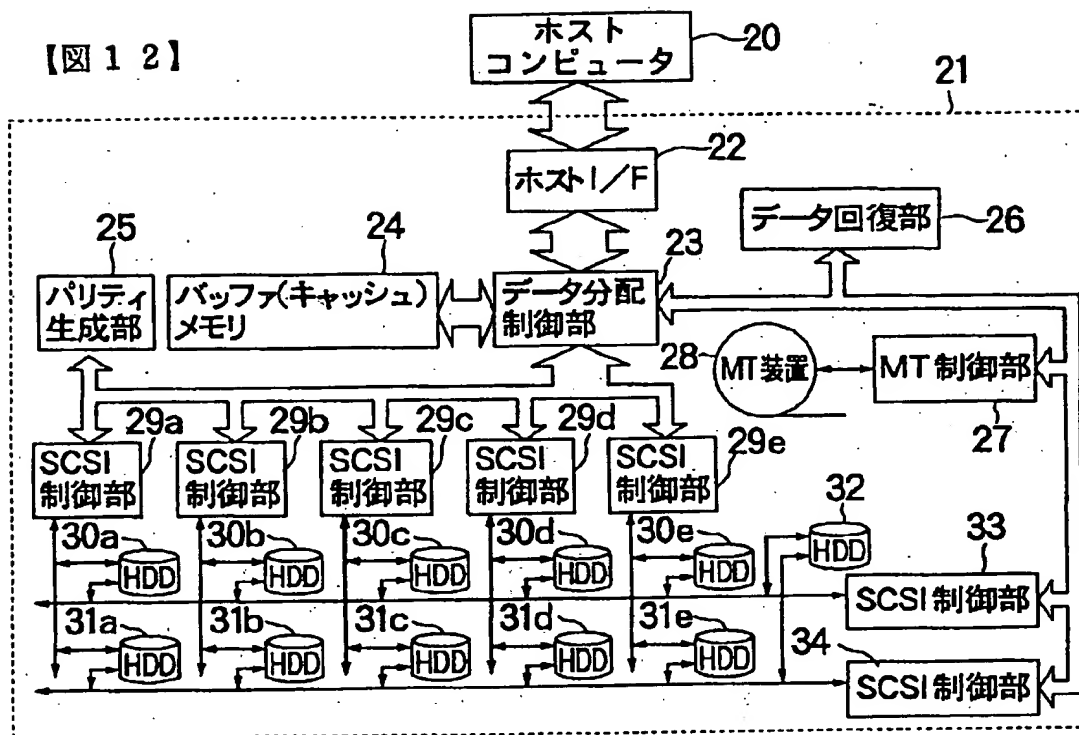


(17)

【図11】

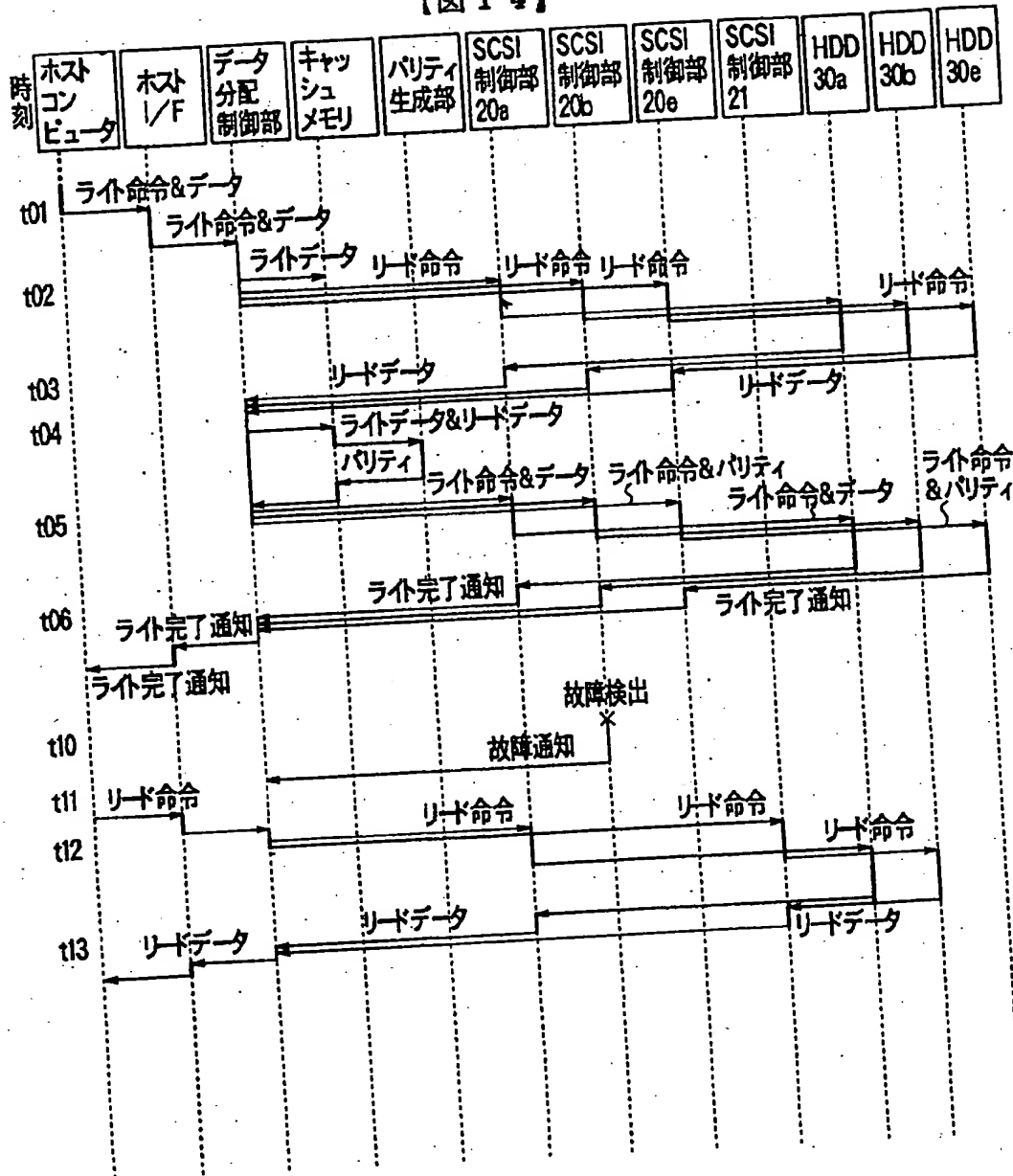


【図12】



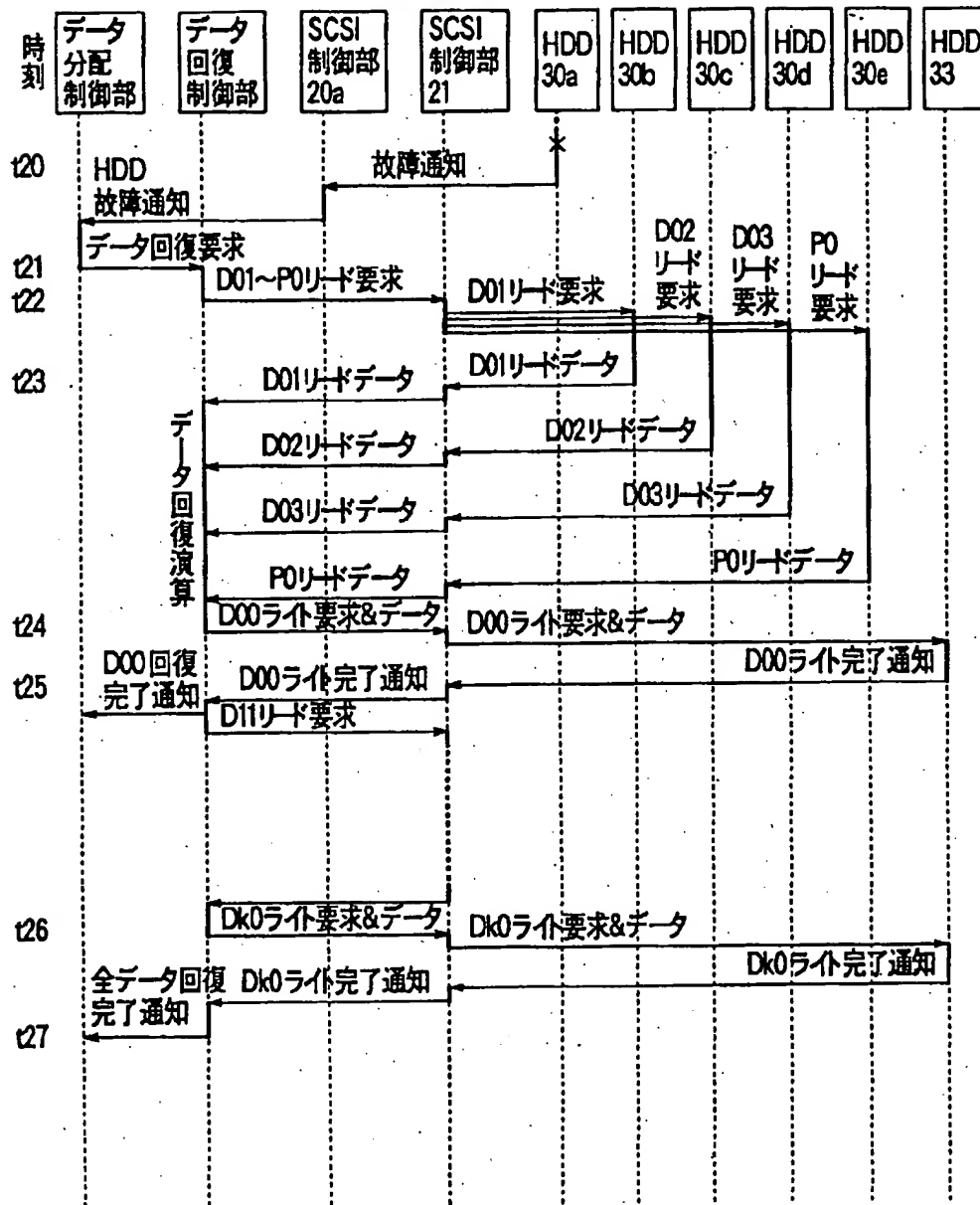
【図14】

【図14】



【図15】

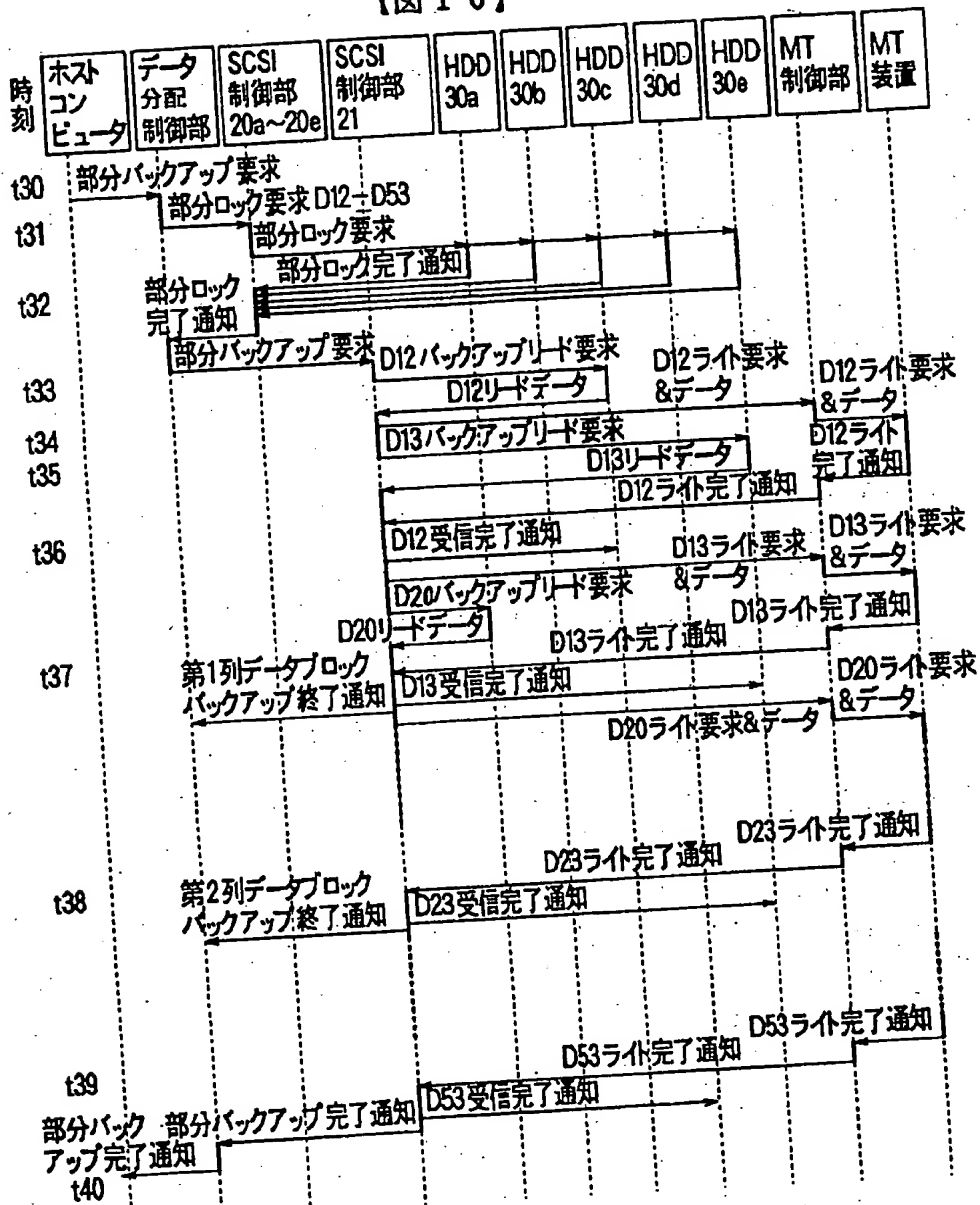
【図15】



(21)

【図16】

【図16】



フロントページの続き

(72)発明者 本田 聖志
 神奈川県横浜市戸塚区吉田町292番地 株
 式会社日立製作所マイクロエレクトロニク
 ス機器開発研究所内

(72)発明者 松並 直人
 神奈川県横浜市戸塚区吉田町292番地 株
 式会社日立製作所マイクロエレクトロニク
 ス機器開発研究所内

(72)発明者 宮沢 章一
神奈川県横浜市戸塚区吉田町292番地 株
式会社日立製作所マイクロエレクトロニク
ス機器開発研究所内

(72)発明者 磯野 聡一
神奈川県横浜市戸塚区吉田町292番地 株
式会社日立製作所マイクロエレクトロニク
ス機器開発研究所内

THIS PAGE BLANK (USPTO)